Enhancing Sketch-Based Image Retrieval by Re-Ranking and Relevance Feedback

Xueming Qian, *Member, IEEE*, Xianglong Tan, Yuting Zhang, Richang Hong, *Member, IEEE*, and Meng Wang, *Member, IEEE*

***Abstract*— A sketch-based image retrieval often needs to optimize the tradeoff between efficiency and precision. Index structures are typically applied to large-scale databases to realize efficient retrievals. However, the performance can be affected by quantization errors. Moreover, the ambiguousness of user-provided examples may also degrade the performance, when compared with traditional image retrieval methods. Sketch-based image retrieval systems that preserve the index structure are challenging. In this paper, we propose an effective sketch-based image retrieval approach with re-ranking and relevance feedback schemes. Our approach makes full use of the semantics in query sketches and the top ranked images of the initial results. We also apply relevance feedback to find more relevant images for the input query sketch. The integration of the two schemes results in mutual benefits and improves the performance of the sketch-based image retrieval.**

***Index Terms*— Sketch, SBIR, relevance feedback, image retrieval, contour matching.**

I. Introduction

METHODS for efficiently searching images are an important research topic. Developments in Internet and mobile devices have increased the demand for powerful and efficient information retrieval tools. Content-based image retrieval (CBIR) mainly uses text and images for queries. However, it is often not possible to precisely describe the content of the desired images using plain text. Additionally, obtaining image examples that exactly match a user's search intentions is not a trivial task. Query sketches drawn by users can effectively describe the aim of a search. Therefore, query-by-sketch is an effective method when text description or query examples are unavailable.

Sketch-based image retrieval (SBIR) methods use a hand-drawn sketch composed of simple strokes or lines to fulfill the image retrieval task [1]. In a user's visual perception, the most informative lines in an image are the contours. A sketch is generally a rough description of an object's shape and contours. The sketch does not need to be artistic, and is simply the user's rough impression of the intended object.

Traditional draw and search systems require that the input sketch is colored and similar to a real photo [3]. This approach converts sketch-based retrieval to content-based image retrieval. The user must draw the sketch carefully and color it to make the sketch visually similar to the natural scene images. Then, CBIR fuses different features (such as shape, color, and texture) together to perform retrieval. However, this method will burden users by requiring detailed drawings, and most importantly, it does not solve the core problem of SBIR, i.e., matching a line-formed sketch and colored images [2].

Image retrieval must deal with the difference between the user's desire and the query example. This difference may be even more severe in sketch-formed queries, because of the ambiguousness in the query sketch caused by a lack of semantic information such as texture attributes [1] and luminance [15]. A simple and similar image is needed for image-based retrieval. But for SBIR, results may vary dramatically if the user's drawing skills are not sophisticated, or if the target cannot be simply depicted using only lines. For example, if a user is looking for pictures of a pyramid but they can only draw a triangle, sketch-based retrieval becomes very challenging [2]. To address this problem, researchers proposed incorporating sketches and text descriptions to disambiguate the input. Lin et al. proposed a method that does not use lines to form the query sketch [4]. The sketch is a drawing that uses different words to represent diverse objects. Their locations and sizes are represented by the words. With the help of these words, the approach first finds some corresponding exemplars, which is then used to search for objects in images. In this sense, it is like a concept-based image retrieval system instead of a sketch-based method.

The problem in sketch-based image retrieval is how to measure the relevance of an image and a query sketch. The similarity measurement can be converted to matching contours and sketches. Effective matching algorithms have received much research attention [2], [5]. Researchers often use global features to match a sketch and an image. The matching algorithm typically uses a predefined tolerance, because the sketches drawn by users are often not precise. However, the global similarity of the sketch and image does not necessarily reflect content similarity. Local feature matching could solve this problem. However, it is computationally intensive, as discussed in [2]. Wang et al., introduced a method that establishes an edge index structure, which

Manuscript received June 11, 2015; revised September 23, 2015; accepted October 13, 2015. Date of publication November 4, 2015; date of current version December 3, 2015. This work was supported in part by the 973 Program under Grant 2012CB316400, in part by the National Natural Science Foundation of China under Project 61173109, Project 60903121, and Project 61332018, and in part by Microsoft Research Asia. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nilanjan Ray. *(Corresponding author: Xueming Qian.)*

X. Qian is with the Key Laboratory for Intelligent Networks and Network Security and the SMILES Laboratory, Ministry of Education, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

X. Tan and Y. Zhang are with the SMILES Laboratory, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: xjtuicemaple@sohu.com; zhangyuting@stu.xjtu.edu.cn).

R. Hong and M. Wang are with the Hefei University of Technology, Hefei 230000, China (e-mail: hongrc.hfut@gmail.com; eric.mengwang@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2497145

1057-7149 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

solves the sketch retrieval problem on large-scale datasets by dramatically reducing the computational cost [2]. They quantize the orientations of each point of the query sketch and the contours in the database into six different angles. Each pixel point of the contours is represented by its orientation and location, referred to as “edgel”. They also proposed an efficient index structure to achieve a fast match.

However, this sketch-based retrieval system heavily relies on the local features [2], and the fault-tolerant rate of the query sketch is comparatively low. Only the images whose shapes are fairly close to the sketch are in the top-ranked list. Moreover, the results also contain noisy images with contours that are partly similar to the query sketch. Some irrelevant images may appear in the top-ranked results. It is important to re-rank the final results and make the top-ranked images more relevant, however this is challenging.

To solve these problems, we propose to optimize the search results at the end of a SBIR system, such as the ARP (angular radial partitioning) [1] or edgel [2], by verifying the top-ranked results and implementing a relevance feedback.

There are a few reasons why an SBIR system typically performs worse than an image-based retrieval system. As previously mentioned, there can be a large difference between a user’s aims and the query example, especially when the query is a drawn sketch. There are also semantic differences between the low-level features and the semantic information of images in the database. These two differences may dramatically degrade the performance.

Relevance feedback has been extensively applied to better interpret users’ search intentions in an interactive way [6]–[9], [38], [39]. It can also be applied to SBIR systems to improve the retrieval performance. However, there are some problems when using relevance feedback in SBIR. There are generally two challenges when applying the relevance feedback technique to SBIR. The first is that the query sketch and returned images do not have the same style. The second is that the scarcity and inaccuracy of a query sketch may mean that many noisy images appear in the top-ranked search results. Thus, we must consider how to select relevant images and get robust feedback.

We propose a system that uses several techniques, including relevant image grouping, re-ranking via visual feature verification (RVFV), and contour-based relevance feedback (CBRF). The aim of grouping approach is to find more relevant images to produce relevant feedback. The RVFV approach removes noisy images and makes the top ranked images more relevant to the input query sketch. The CBRF approach uses the contours of the top-ranked images obtained by the SBIR system as new queries to find more relevant images. We apply RVFV again to remove irrelevant images that introduced in the CBRF stage. The two systems are both offline and are considerable enhancements on SBIR. With a small increase in complexity, the sketch retrieval system can retrieve more desired images.

The main contributions of this paper are summarized as follows. 1) We propose an effective sketch-based image retrieval approach with relevant image grouping, verification and re-ranking. The semantics explored from the sketch and the local features of the verified relevant images are fused to reduce the user’s search intention gap in SBIR. 2) We propose mining relevant images in the top-ranked results from the initial SBIR system using relevant image grouping, and using them in the relevance feedback. 3) We propose a visual verification system that re-ranks the results to improve the overall performance. 4) We integrate a contour-based relevance feedback system into the SBIR system to improve the retrieval performance. This method uses contours as sketches to carry out the relevance feedback in SBIR. We test our relevance feedback based SBIR approach on the ARP and edgel based SBIR systems. The results demonstrate that we have achieved improvements with very little increase in the computational cost.

The remainder of this paper is organized as follows. Work related to sketch-based retrieval is reviewed in Section II. We describe the proposed approach in Section III, our experiments in Section IV, and the discussions in Section V. Finally, we present our conclusions in Section VI.

II. Related Work

Many SBIR methods have been proposed over the past 20 years. Query by visual example [10] defines a pictorial index for each image, and computes the correlations between the corresponding indexes to retrieve the results. An image is divided into equalized blocks and the correlation is calculated by shifting these blocks.

Zernike moment is a moment invariant method that has been used in SBIR [12], [13]. It can solve the rotation, scale, and translation invariant problems. The method in [13] uses Zernike orthogonal polynomials to extract the Zernike moment descriptor of an image, and uses the Manhattan distance to measure the similarity between a sketch and image. The edge histogram descriptor (EHD) and the histogram of oriented gradients (HOG) are also used to establish the SBIR system [14]. They are both global features extracted from the edges of images. Chalechale et al. proposed an angular partition approach that divides the edge into several blocks in terms of orientations [41]. An angular radial partitioning-based SBIR approach was proposed in [1], which considers the radial factor during the retrieval process.

Most existing methods mainly use global features or divide images into blocks to represent the image [12]–[15]. These methods do not work well because of the ambiguousness of sketches and shapes. Additionally, the incompleteness of a user’s drawing may also affect the results. Consequently, researchers proposed exploring the local saliency in SBIR. Chen et al. used a freehand sketch and some text labels to search for Internet images [16]. Although this method was very accurate, it was very computationally expensive. Thus, a SBIR with index structures is more appropriate for a large-scale image set, and achieves the best balance between the retrieval performance, and time and storage costs.

The edgel index approach is a shape-based indexing method [2]. It solves the shape-to-image matching problem using pixel level matching. Oriented chamfer matching [17] is used to compute the distance between contours.

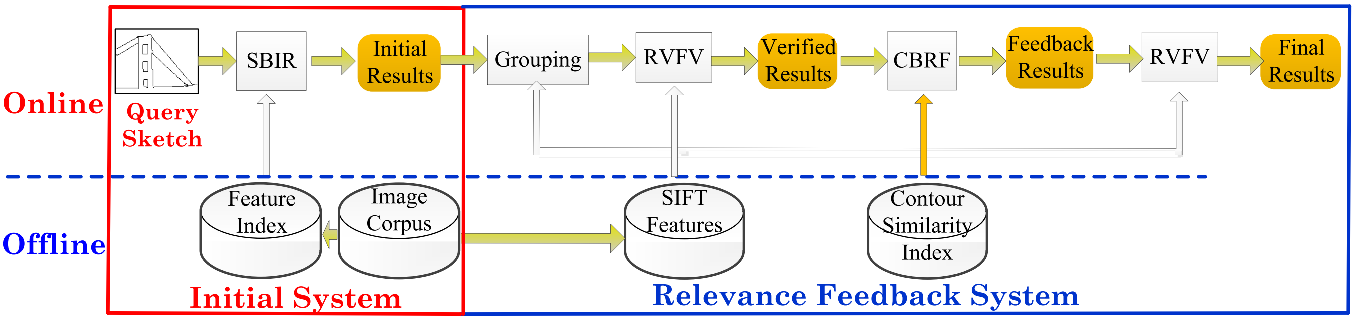


Fig. 1. The framework of our system.

To conveniently build the index structure, Wang et al. used a binary similarity map (a hit map) instead of the distance map [2]. For each input sketch, N hit maps are created, which correspond to the N orientations. They also designed a simple hit function. Specifically, if a point falls in the valid region on a hit map in the same channel, it is considered as one hit. The sum of all the hits is the similarity between the image D (represented by its contours) and the query sketch Q. Then, they build an edgel index structure for fast retrieval.

Wang et al. also proposed a two-way matching method [2]. They computed the similarity between D and Q and the counterpart from Q to D. Then, they multiplied the two similarity scores to obtain a final score that reduces the influence of trivial results. To avoid edgel index distortion, they simply choose the top N candidate images for the Q to D process. For images with salient structures, they also proposed a structure consistent sketch matching approach. It decomposes one query sketch into multiple sub-queries, and uses the geometric means of similarity scores from all these sub-queries as the final score.

The ARP [1]-based SBIR approach refines the angular partitioning (AP) feature [41] using radial partitioning (RP). The ARP feature is obtained by partitioning the edge image $I(\rho,\theta)$ into A × B sectors. It uses the image center as the center of the circles. A is the number of angular partitions and B is the number of radius partitions. The range of each angle is $\theta = 2\pi/A$ and the radius of successive concentric circles is $\rho = R/B$, where R is the radius of the surrounding circle of the image. Based on the edge obtained from the original image $I(\rho,\theta)$, each sector is represented by its corresponding edge pixel number.

Relevance feedback is applied in our method to return more relevant images. Generally speaking, there are three types of feedbacks: explicit feedback [7], implicit feedback [8], and blind feedback [9], [38], [39]. Explicit and implicit feedbacks require interaction between users and the system. Explicit feedback directly obtains relevance judgments, whereas implicit feedback is inferred from a user's behavior. However, this burdens users and induces certain delays (because the system must wait for a user response) [39]. Blind relevance feedback, also known as pseudo relevance feedback [6], applies feedback automatically without interactions with users. It carries out normal retrievals and assumes that some of the top N results are relevant. This method is more appropriate because of its timely response, so it is extensively used in retrieval methods [18]–[20]. Multimodal contextual information has been shown to more effectively re-rank image results [38], [39].

Structural relations and global shape descriptors have been used to represent the content of sketches, and for relevance feedback based on a biased SVM (support vector machine) [11]. An explicit relevance feedback method was used in [11], which refined the results based on user interactions. The main differences between the method in [11] and the proposed method are: 1) we use blind feedback (which does not need user interactions) whereas [11] used explicit feedback (which is hard to apply and is inconvenient for users); and 2) our retrieval approach uses efficient indexing, whereas [11] trains classifiers.

III. Sketch-Based Image Retrieval With Relevance Feedback

The framework of the proposed SBIR system is shown in Fig. 1. It consists of two parts: the offline part and the online part. Our approach can be included at the back end of any initial SBIR system (such as the edgel [2] and ARP [1] methods) using relevance feedback to improve performance. We now focus on an edgel SBIR system to illustrate our approach. In the offline part of the method, we must build an edgel index structure for each image based on the Berkeley edge detector [21]. Then, we extract SIFT features and record the SIFT descriptors with their locations and orientations. Finally, we build a contour similarity index for each image.

In the online part, for a given input query sketch, we sequentially execute five stages: 1) the initial SBIR [2], which obtains the initial result shown to the left of Fig. 1; 2) relevant image grouping for the initial results, which finds the relevant images from the top R images in the top N ranked results; 3) re-rank and verify the results using SIFT matching; 4) contour-based relevant feedback to find more relevant images; and 5) re-rank the results of the relevant feedback to improve the performance.

A. Sketch-Based Image Retrieval

In the offline system, we build a feature index structure (such as edgel) for each image, as in [2]. More details can be found in [2]. We give a brief overview of the approach, which consists of the following three steps.

1) For an image database with T images, we apply the Berkeley detector [21] to each image (resized to 200×200). This produces hit maps with six orientation

channels ($\theta = 6$). Thus, for each image, we build an index structure with 200×200×6 entries for the six orientation channels.

2) The Berkeley detector [21] extracts contours. It uses the brightness, color, and texture gradients to accurately detect and localize the boundaries of images.
3) For each point at a certain orientation, we build an inverted list for fast indexing (i.e., the edgel index structure used in [2]). For each edgel point in the contours, the position (x,y) and quantized orientation channel θ are combined to (x,y,θ). For each entry (x,y,θ), we build an inverted list of images (IDs).
4) When a query sketch Q (normalized to 200×200 entries) is input to the system, six hit maps are generated by marking the regions surrounding the sketch lines within a certain radius, and quantizing each edge orientation into six channels [2]. By comparing the edgels (x,y,θ) of the hit maps of the query sketch and the edges extracted from the database images, we can measure the similarities between the sketch and images. Each edgel marked in the hit maps is used to search the inverted list for corresponding image IDs. Finally, the similarity between the query sketch (Q) and the image (D) in the database is computed by counting how many times D appears during the search.

We sort the similarity scores in descending order, and determine the initial results (the N top ranked). In the following steps, we apply re-ranking and relevance feedback schemes to these N images.

B. Relevant Images Grouping for Relevant Feedback

The top-ranked images obtained by the initial SBIR may contain irrelevant images. In our approach, the relevant images are the ones that occur most in the top N images. We make full use of the top R images ($R < N$) to find relevant images for CBRF. Our approach is motivated by retrieval results clustering, which improves the diversity of top-ranked results [42], [43] by finding near duplicated image groups [44]–[47].

We apply near-duplicate image clustering to the top ranked R images to find similar images from the top N initial SBIR results [46]. This approach consists of the following steps.

1) For each image, we record the SIFT descriptors together with their locations (x,y) and orientations [30], [31]. The SIFT feature extraction is carried out off-line for the dataset images.
2) We first find near-duplicated images for the top R images of the top N images returned by the initial SBIR, as shown in Fig. 1. We use the similarity measurement (i.e., near-duplicate image detection) with the existing image matching approach [37], [46]. In this paper, we use binary edge-SIFT to carry out the near-duplicate image retrieval approach and find near-duplicate image groups.
3) We further cluster the detected near-duplicate images into groups for the top ranked R images. Assume that the group number is K ($K \leq R$) and we record the corresponding image numbers.
4) We use the cluster with the most near duplicate images as relevant image group for the query sketch. At the same time, we set the initial scores of images in the relevant image group as their maximum, and the initial scores of the irrelevant images as their minimum. This step ranks the images in the relevant image group ahead of the other images.

Using relevant image grouping, we can roughly eliminate the noisy images from the top-ranked results. Then, we further use the top N images with RVFV to obtain more relevant images.

We use the duplicate image group from the top R-ranked images (denoted by top-R+top-N), rather than the top N images to eliminate noise. Generally, a higher-ranked image is more relevant to the query sketch. If we use the top-ranked N images directly in RVFV, we will include some noise. This would negatively impact the final CBRF. More discussions are given in our experiments.

C. Re-Ranking via Visual Feature Verification

Although the relevant image grouping approach can find more relevant images for the query sketch, some irrelevant images may appear in the top N results. If we re-rank the top N results by measuring their similarities in the visual feature space, then the refined search results will be more satisfactory.

Our aim is to filter out irrelevant images using content matching or spatial constraints [22], [23], [37], [47], which are often used in retrieval result verifications [22]–[30]. Thus, in this paper, we leverage the advantages of both retrieval result verification and relevance feedback to improve the retrieval performance.

We apply RVFV twice, as shown in Fig. 1. The first time reduces the number of false positive results, and the last time optimizes the final results. RVFV consists of two steps: 1) finding SIFT pairs of the standard image and other images; and 2) re-ranking using the similarity scores.

1) Feature Matching: In this paper, RVFV is only applied to the top N initial results. We select some of the relevant images from the top N-ranked images to expand the query and get more relevant results. We find SIFT pairs of the standard image (the top-ranked image after relevant image grouping of the initial SBIR results, I_S) and other images (the top-ranked N images, but not including duplicates of the standard image).

The similarity scores are measured using matched SIFT point pairs. P_A is a SIFT point in image I_A, and P_B is a SIFT point in image I_B. We define $(P_A P_B)$ as a SIFT pair, if and only if, the best-matched SIFT point of P_A of image I_A in image I_B is P_B, and vice versa.

The similarity of two SIFT descriptors (d_1 and d_2) is measured using the L_2-norm [32]. That is,

$$\begin{aligned}(d_1, d_2) &= ||d_1 - d_2||_2^2 = \sum\nolimits_i |d_1^i - d_2^i|^2 = \sum\nolimits_{i|d_2^i=00} |d_1^i|^2 \\ &\quad + \sum\nolimits_{i|d_1^i=00} |d_2^i|^2 + \sum\nolimits_{i|d_1^i \neq 0, d_2^i \neq 0} |d_1^i - d_2^i|^2 \\ &= ||d_1||_2^2 + ||d_2||_2^2 \\ &\quad + \sum\nolimits_{i|d_1^i \neq 0, d_{2i} \neq 0} (|d_1^i - d_2^i|^2 - |d_1^i|^2 - |d_2^i|^2) \\ &= 2 - \sum\nolimits_{i|d_1^i \neq 0, d_2^i \neq 0} d_1^i d_2^i, \end{aligned} \tag{1}$$

where d_*^i is the value of d_* in the i-th dimension, for $i = 1, \dots, 128$. d_*^i is normalized using

$$d_*^i = d_*^i / ||d_*||_2^2. \tag{2}$$

Thus, in (1), we have $||d_1||_2^2 + ||d_2||_2^2 = 2$.

According to [32], the similarity score between d_{Ai} of image I_A and d_{Bj} of image I_B is defined as

$$\underset{d}{Sim}(d_{Ai}, d_{Bj}) = \sum\nolimits_{l|d_{Ai}^l \neq 0, d_{Bj}^l \neq 0} d_{Ai}^l d_{Bj}^l, \tag{3}$$

where d^l denotes the value of the l-th dimension of the descriptor d.

Based on (3), the similarity score is

$$\underset{d}{Sim}_N(d_{Ai}, d_{Bj}) = \frac{\underset{d}{Sim}^2(d_{Ai}, d_{Bj})}{\frac{1}{L_B}\sum_{k=1}^{L_B} \underset{d}{Sim}(d_{Ai}, d_{Bk}) * \frac{1}{L_A}\sum_{k=1}^{L_A} \underset{d}{Sim}(d_{Ak}, d_{B}j)}, \tag{4}$$

where L_A and L_B are the number of SIFT points in image I_A and I_B, respectively. The denominator serves as a normalization, considering the average similarity between d_{Ai} and all other descriptors in image I_B, and the average similarity between d_{Bj} and all other descriptors in image I_A.

2) Similarity-Based Re-Ranking: SIFT feature matching has been extensively applied to image classification [30], [33]–[35]. Considering the spatial locations, orientation, or other geometric constraints [36], [37] can improve matching performances. Sketch-based image retrieval has strong spatial constraints.

Therefore, we use SIFT locations (L) and orientations (O) to add weights to matched SIFT pairs. The weight is defined as

$$W(m) = \exp(-\alpha \times (W_L(m) + \beta \times W_O(m)), \tag{5}$$

where m denotes the m-th SIFT pair between I_A and I_B. α controls the convergence of the exponential function, and β balances the two parts. $W_L(m)$ and $W_O(m)$ are the location and orientation weights, respectively. They are defined as

$$W_L(m) = ||L(A_m) - L(B_m)||_2^2 \tag{6}$$

and

$$W_O(m) = \min(|O(A_m) - O(B_m)|, |O(A_m) + O(B_m)|), \tag{7}$$

where $L(.)$ and $O(.)$ are the location and orientation of a SIFT point, and (A_m, B_m) is the m-th SIFT pair of I_A and I_B. We use the minimum of the difference and the sum of orientations so that $W_O(m)$ is in the range $[-\pi, \pi]$.

Then, the similarity between two images can be determined by summing the weighted scores of the matched SIFT point pairs. That is,

$$SIM(I_A, I_B) = \sum_m \underset{d}{Sim}(d_{Am}, d_{Bm}) W(m). \tag{8}$$

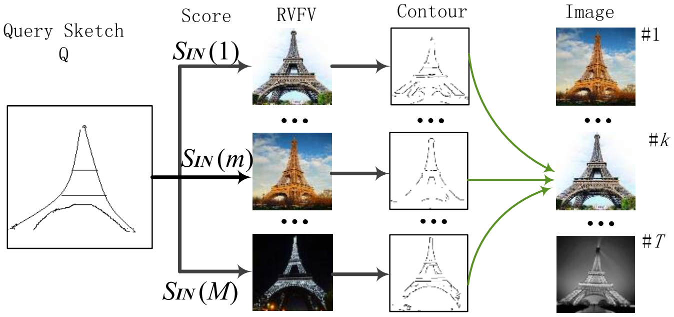


Fig. 2. Proposed contour-based relevance feedback system. For a given query sketch, the initial results are obtained using the SBIR system and M verified images remain after the RVFV system. Then, the contours of M images in the verified results are used as new queries to search for more similar images, and the final results are determined by combining the scores from the two stages.

For the top N results of the initial retrieval ($N = 100$ in our experiments), we compute the similarity of image I_k to the standard image I_S using

$$S_k = SIM(I_s, I_k), \quad k = 1 \sim N. \tag{9}$$

When $k = 1$, we have $S_k = 1$. S_k indicates how similar an image in the initial result is to the standard image. We evaluate if it satisfies a minimum matching requirement (i.e., S_k is larger than a cut-off threshold), or we sort S_k in descending order and select the top M images. The selected images are used for the contour-based relevance feedback.

D. Contour-Based Relevance Feedback

It is useful to expand the query for image-based retrieval to improve the final result [22]. A sketch is a description of contours. The contour of a top-ranked image can also be regarded as a sketch and used to return more relevant images.

Our relevance feedback algorithm contains the following steps.

1) The contours of the verified images are used as new query sketches.
2) Each image in the corpus is given a score based on each of the new query contours.
3) The final similarity score of each image in the corpus is obtained by combining the scores of the initial and expanded retrievals.
4) The final ranked list is generated using the initial system for each new query. These ranked lists are combined and used to add weight to the initial result and obtain the final ranked list.

Assume that M relevant images are obtained through the first RVFV ($N \geq M$). Then, CBRF finds more relevant images using the contours of the M images as new query sketches. After the above query expansion, we get ranked lists for the M-expanded query sketches. We compute the relevance feedback scores of each image in the corpus for each expanded query sketch, as shown in Fig. 2. The corresponding images and scores for the M contours are determined using the contour similarity index structure.

An image I_k in the corpus has M scores after the query expansion. Its feedback score should be the weighted sum of these M scores. Accordingly, the relevance feedback score $S_{RF}(k)$ is

$$S_{RF}(k)=\sum\nolimits_{m=1}^{M} S_{ID}(m,k)\times S_{IN}(m);\quad k=1,\cdots,T, \tag{10}$$

where $S_{ID}(m,k)$ is the score of the image I_c for the m-th expanded query after the first RVFV. $S_{IN}(m)$ and $S_{RF}(k)$ are the scores of the initial retrieval and relevance feedback, respectively. T is the total number of images in the image corpus.

A higher-ranked image in the initial results has more influence on the feedback. We make full use of this to generate the final ranked retrieval results and ensure the feedback is positive. The score of image I_c is

$$S(k)=(1-w)\times S_{IN}(k)+w\times S_{RF}(k);\quad k=1,\cdots,T, \tag{11}$$

where w is the feedback weight that determines the importance of the feedback and is in the range [0,1]. In our baseline approach, we set $w = 0.4$. More discussion regarding the influence of w on the final performance is given in Section V.

After CBRF, we have a new ranked list. It is likely that some irrelevant results are introduced by the M expanded queries. So we apply RVFV again to re-rank the top N results. Among the top N results, some images are new and some have already been verified. Images verified in the first RVFV are recorded so that we do not need to recompute their scores, so the second RVFV is much faster than the first.

IV. Experiments

To demonstrate the effectiveness of the proposed approach, we compared our algorithm with the edgel [2] and ARP [1] methods on the existing SBIR_100k dataset and our own dataset. We used Matlab on Windows 7 to calculate the SBIR and relevance feedback, and Berkeley Detector [21] implemented using Matlab on Ubuntu to extract the contours.

A. Datasets

1) SBIR_100K Dataset: This dataset was used in [40] (denoted as dataset_100k) and contains 101,240 images. There are 1240 benchmarked images for 31 query sketches, and 100,000 noise images.

2) Our Dataset: Our dataset consists of 296,562 images. It contains a sketch-describable dataset of 68,647 images gathered from Google using keywords to search for relevant images. The search results in Google were collected to form the image corpus. Examples of keywords are apple, hat, shoes, bike, cat, tower, notebook, airplane, car, and pyramid. These keywords were carefully chosen so that the collected images were unambiguous and could be easily described by user sketches. Topics mainly included living goods, fruits, animals, and landmarks that can be easily sketched. There were approximately 1000 images in each topic. This dataset also contains the GOLD set [33], [34], [48], which mainly contains landmarks and landscapes. It mostly contained images with different topics to the images gathered from Google.

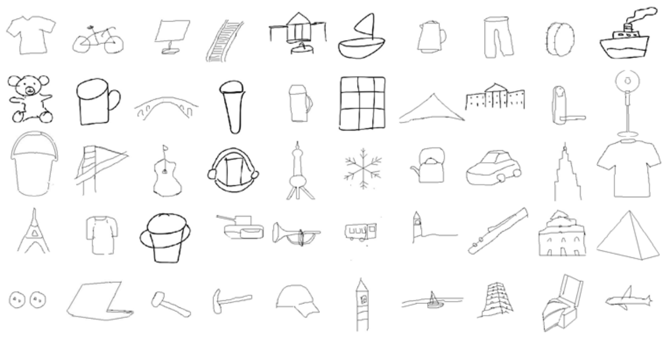


Fig. 3. Example sketches.

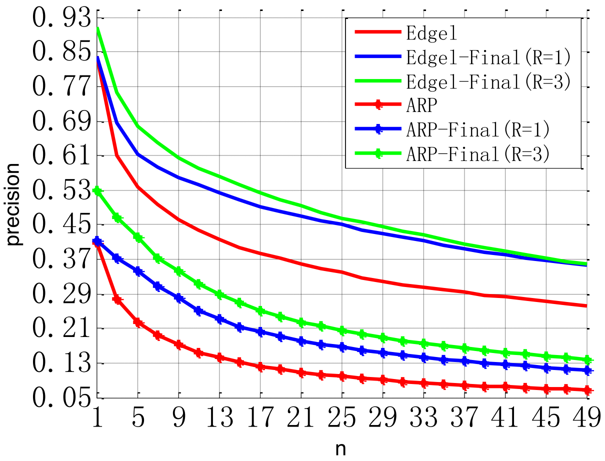


Fig. 4. Performances using our dataset.

We drew 361 query sketches, including 162 good sketches drawn by 10 students with excellent drawing skills and 199 inferior sketches drawn by students in our lab. Some of them are shown in Fig. 3.

B. Performance Evaluation

We used the precision under depth n (denoted as *Precion@n*) to measure the objective performance, defined as

$$Precion@n=\frac{1}{Z}\sum\nolimits_{t=1}^{Z}\frac{1}{n}\sum\nolimits_{i=1}^{n}R_t(i), \tag{12}$$

where $R_t(i)$ is the relevance of the i-th result for query t, $i \in [1,2,\ldots,n]$, and $t \in [1,2,\ldots Z]$. $Z = 361$ for our dataset, and $Z = 31$ for the SBIR_100K dataset. If it is relevant to the query sketch then $R_t(i) = 1$, otherwise $R_t(i) = 0$.

C. Objective Comparisons

The *Precision@n* curves of the proposed approach on the initial SBIR system based on the edgel and ARP methods for depths in the range of [1,50] are shown in Fig. 4 (our dataset) and Fig. 5 (SBIR_100K dataset). Our method used the SM (SIFT matching), LW (location weighting), and OW (orientation weighting) schemes in the first and second RVFV,

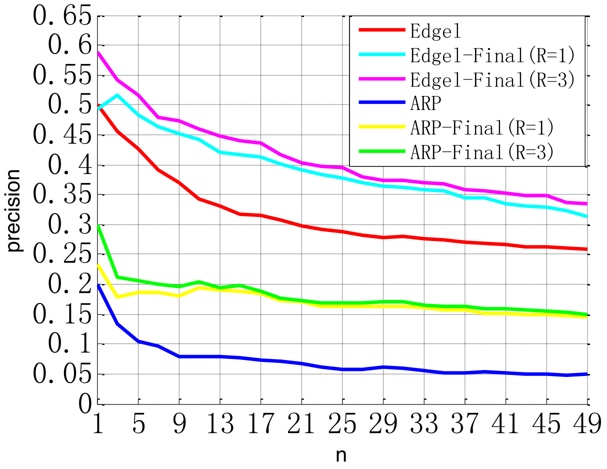


Fig. 5. Performances using the SBIR_100K dataset.

TABLE I

COMPUTATIONAL COSTS OF EACH STEP OF THE PROPOSED METHOD COMPARED WITH THE INITIAL SBIR APPROACHES USING THE EDGEL AND ARP METHODS (IN SECONDS)

	Initial	ours				
	SBIR	Clustering	RVFV1	CBRF	RVFV2	Total
Edgel	9.77	0.017	0.73	0.14	0.41	11.06
ARP	0.64	0.015	0.53	0.10	0.26	1.55

and used CBRF. In the RVFV procedure, we set $\alpha = 0.05$, $\beta = 50$, $N = 100$, and $M = 50$. In the CBRF procedure, we set $w = 0.4$ (the weight of the feedback score).

In the Fig.4 we compared our approach to the edgel and ARP methods with $R = 1$ and $R = 3$, denoted by "Edgel-Final ($R = 1$)" and "Edgel-Final ($R = 3$)", and "ARP-Final ($R = 1$)" and "ARP-Final ($R = 3$)".

From Fig.4, we found that the proposed algorithm was 10% more accurate than the edgel [2] and ARP [1] methods for the top 10 results when $R = 3$. For $n = 1$, our method performed as well as the existing approaches when $R = 1$, because we used the top ranked image as the standard image for the verification and relevance feedback procedures. When $R = 1$, the proposed algorithm was 10% more accurate than the edgel and ARP methods for the range from top-5 to top-30 results. From Fig.4, we find that the relevant images grouping play an important role to improve the accurate for the top-1 result.

From Fig.5, we can find the proposed algorithm has the similar situation with Fig.4. The difference is the decline slowed. The reason is that in SBIR_100K dataset, the initial results contain more relevant images and the proposed method can find them at the top ranked results.

To measure the computational cost of our algorithm, we applied the method to the 361 queries. The average computational costs of the three methods are shown in Table I. These experiments were implemented using Matlab on Linux, and the code was only optimized in Matlab. Therefore, the computational cost was more than reported in [2]. But the relative computational costs were obviously different to those of the initial SBIR method. The average computational cost of the edgel method was 9.77 s. The computational costs of the first RVFV (including relevant image grouping), CBRF, and the second RVFV stages of our method were 0.73 s, 0.14 s, and 0.41 s, respectively. The total time taken by our relevance feedback system was 1.28 s, which was less than 1/7th of the time taken by the edgel method. For the ARP method, our system took 0.91 s to calculate the relevance feedback.

TABLE II

THE VALUE OF AP (K)% FOR DIFFERENT α AND β

β / α	0.1	1	10	20	50	100	1000
0.01	45.63	45.68	46.00	46.33	46.94	47.56	47.81
0.05	49.26	49.33	49.77	49.91	49.96	49.71	46.95
0.1	49.54	49.60	49.92	49.92	49.62	49.15	46.40
0.5	47.28	47.56	48.01	48.05	47.97	47.81	45.99
1	46.54	46.93	47.60	47.73	47.69	47.62	45.98
5	46.16	46.59	47.42	47.51	47.56	47.52	45.93
10	46.10	46.56	47.41	47.48	47.55	47.51	45.57
50	45.85	46.27	46.83	46.75	46.36	45.88	43.92

V. DISCUSSIONS

We now discuss the impacts of the parameters on the performance of our sketch-based retrieval system. For a fair and clear comparison, we used the good sketches as inputs. We set $R = 1$. We considered the weights α and β (which compute the similarity score in RVFV), M (which selects relevant images from the ranked list after RVFV), and the weight w in the CBRF system (which computes the similarity scores). We compared different verification approaches in RVFV, and the contributions of the different procedures on the entire system. After this, the parameters R and N in the relevant image grouping process were investigated. Finally, we considered the impact of the quality of the input sketches.

A. Impacts of α and β in RVFV

There are two parameters in RVFV, α and β. They determine the variation in similarity scores due to the combination of the location and orientation differences. Because the images are all resized to 200×200, most location distances between suitable pairs are between 0 and 100. So α should be in the range of 0.01 to 0.2, to ensure that the exponential function converges.

The orientation difference ranges between 0 and 2π, and β should be in the range of 20–50 so that the two factors have the same order of magnitude. The impacts of these parameters on the final SBIR are shown in Table II. We used an $AP(K)$ index, which is the average of all points in *Precision@n* curve. That is,

$$AP(K) = \frac{1}{K}\sum\nolimits_{n=1}^{K} Precison@n. \qquad (13)$$

Table II shows that the best performance was achieved when $\alpha = 0.05$ and $\beta = 50$.

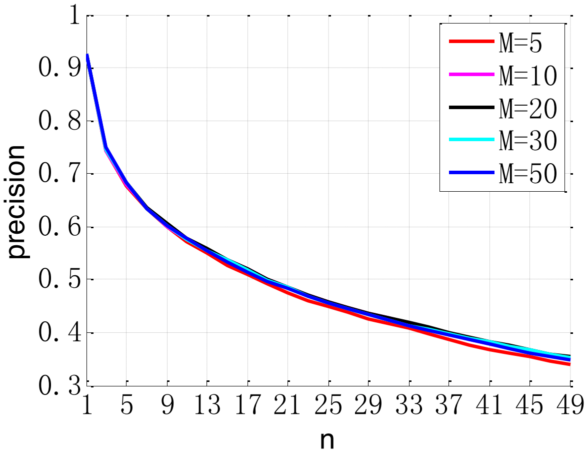


Fig. 6. The effect of M on the proposed method.

B. Impact of M in RVFV

After the first RVFV, the relevant images are selected to ensure the input to CBRF is valid. We select the top-ranked M images.

In our baseline approach, we set $M = 50$. The performances of our approach when M varied in the range [5, 50] are shown in Fig. 6. T The best performance was achieved when $M = 30$, but the performance did not significantly vary for different M.

C. Impact of w in CBRF

The parameter w in (11) is used to compute the scores for the relevance feedback. We set $w = 0.4$ in our baseline experiments. This parameter determines the contributions of the initial and relevant feedback systems. Accordingly, w should range between 0 and 1. $w = 0$ means that no CBRF will be applied. $w = 1$ means that the relevance feedback scores determine the final ranked list and the initial results are discarded. As shown in Fig. 7, the method performed best when w was approximately 0.3. From Fig. 7, we find that the initial retrieval results are important to the CBRF. The CBRF did not significantly improve the final performance for the following reasons.

1) The CBRF is also based on the top-ranked results of the SBIR. We use the contours of the top ranked images as input to search for relevant images. From this point of view, the initial SBIR and CBRF are similar.
2) CBRF introduces some relevant images, but also some irrelevant images. Applying RVFV to the CBRF removes some irrelevant images and improves the performance.

D. Impact of Different Matching Approaches in RVFV

Bag-of-words (BOW) features and inverted file list schemes are widely used in retrieval [32], [35], [36]. In the proposed RVFV, we match images using the original SIFT descriptors instead of the BOW features. There are two general reasons for this. First, the RVFV scheme is only applied to the top-ranked images, which is not computationally intensive. 2) BOW is obtained using quantization, which introduces uncertainties that degrade the matching performance.

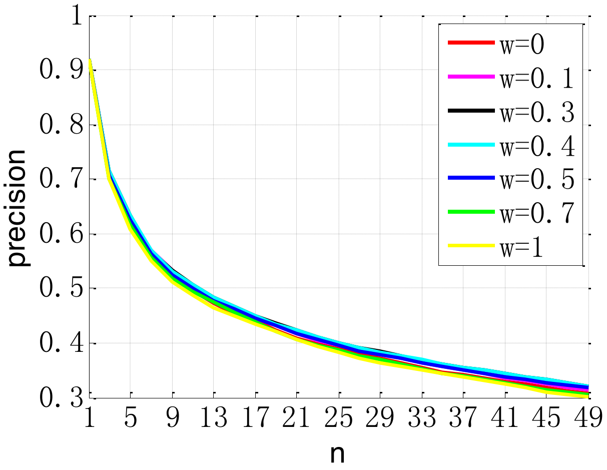


Fig. 7. Precision@n curves for various w.

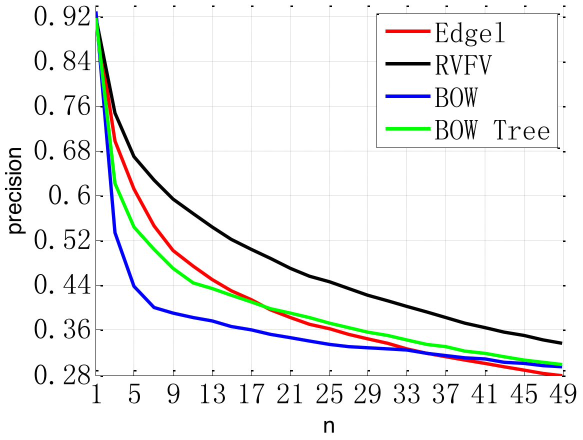


Fig. 8. Comparison of our RVFV and BOW schemes.

Thus, we used the SIFT descriptor matching scheme with location and orientation weighting. We compared the following four methods.

1) The edgel [2] method without RVFV.
2) The RVFV procedure using weighted SIFT pairs (see Section III C).
3) Matching with BOW histograms (denoted by BOW). The similarity of two images is calculated using the Euclidean distance between their BOW histograms.
4) Matching with BOW tree structures (denoted by BOW tree). A vocabulary tree (a hierarchal inverted list index structure) is established and uses the TF-IDF scheme to determine weights [36].

For the methods in 3) and 4), we used two BOW based methods to replace the matching scheme in RVFV. SIFT features of images in the database were quantized into visual words using an eight-level hierarchal quantization scheme [36]. The total codebook contained 1,600,000 entries.

The performances of the four approaches are shown in Fig. 8. We can conclude that only RVFV was stable. This illustrates that strong constraints produces valid retrieval result verifications.

As discussed in Section III C, many SIFT matching approaches simply use the 128-dims SIFT descriptors to measure the similarity between images. We also use the location

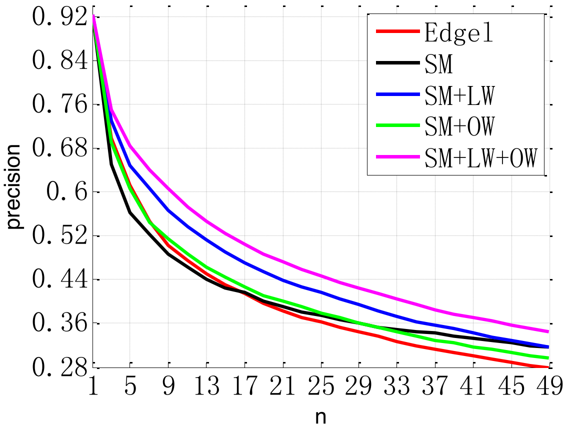


Fig. 9. Comparison of different verification methods in RVFV.

and orientation of SIFT points in the weighting scheme. This is because the initial SBIR system is a pixel-level matching system [2], which is sensitive to changes to the location and orientation. However, SIFT features focus on scale-invariance without any restriction on location and orientation. So the SIFT matching algorithm applied to the results after grouping can be enhanced if location and orientation constraints are introduced. We should ensure that the points of every SIFT point pair have similar SIFT descriptors, locations, and orientations. This reduces the impact of false matches, and improves the performance.

More results for the weighting schemes in RVFV are shown in Fig. 9. We systematically compared the performances of SM, SM+LW, SM+OW, and SM+LW+OW with the edgel method, where SM refers to re-ranking with original SIFT matching, LW refers to location weighting, and OW refers to orientation weighting. Figure 9 demonstrates that the re-ranking scheme may be useless when only original SIFT matching is applied. With the help of LW and OW, the results after re-ranking can outperform the initial results. From Fig. 9, we can see that the location information was important in RVFV, but the orientation weighting also had a positive contribution. The method performed better when we used both location and orientation weighting.

E. Contributions of the Different Parts of Our Approach

As shown in Fig. 1, our method mainly consists of the first RVFV, CBRF, and the second RVFV. Figure 10 shows the contributions of each part. The performance obviously improves after the RVFV. The CBRF curve in Fig. 10 shows that it cannot directly enhance the relevance feedback system. In our experiments, we found that the second RVFV cannot further improve the performance without CBRF. The CBRF introduces more relevant images, and the final performance (denoted by final) is more reasonable after the second RVFV. This figure shows that the different parts of our system work well and all improve the performance to some extent.

F. Impact of Relevant Image Grouping

As shown in Fig. 1, our CBRF-based SBIR approach mines the relevant images for the CBRF from the top-ranked results. We use near-duplicate image clustering for the top R-ranked images to find the most near duplicated image groups in the top N (R$<N$) initial SBIR results. N determines how many images in the top list of initial results should be involved in the RVFV procedure. Theoretically, N should be as large as possible to include more images in the re-ranking procedure. However, N does not need to be very large, because relevant images will probably have higher ranks.

In Part B of Section III, we used the top R results to find groups of relevant images, rather than directly using the top N. Alternatively, we can use the following two approaches. We can determine relevant images by grouping the top R and using the top-ranked image as the standard image. Then, we can use the top N results for RVFV (denoted by Top-R+Top-N). In the other approach, we directly determine the relevant image groups from the top N-ranked initial results (denoted by Top-N), and use these as the input to RVFV. The performances of these two methods are shown in Fig. 11, Fig. 12, and Fig. 13. Figure 11 shows the performances of our approaches compared with the edgel method with R={1,3} and N={50,100}. In Fig. 12, we compared the performances of our approaches to ARP with R={1,3} and N={50,100}.

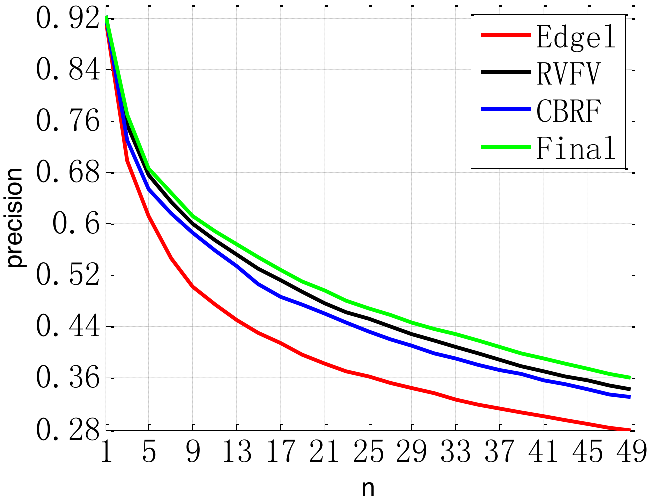


Fig. 10. The contributions of the different procedures.

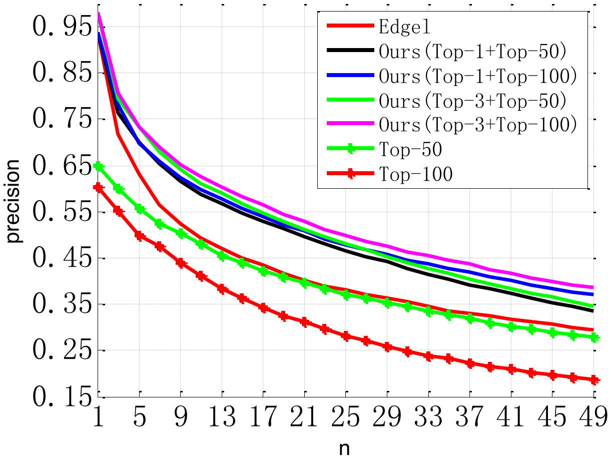


Fig. 11. Impacts of the relevant image grouping on the edgel method.

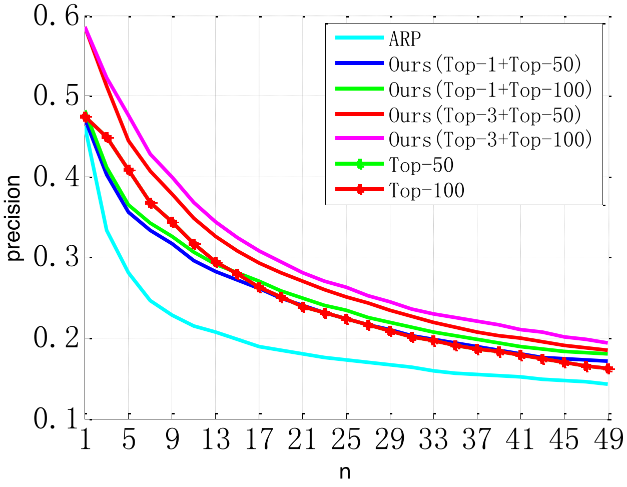


Fig. 12. Impacts of the relevant image grouping for ARP.

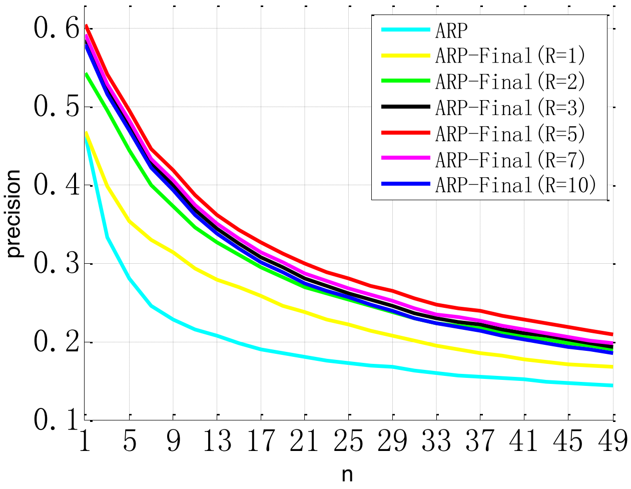


Fig. 13. Impacts of the relevant image grouping for ARP, for different values of R.

For the same R, the proposed approach compared to the edgel and ARP methods performed better for larger N. This shows that finding relevant image groups from more sources improves the performance.

Moreover, the Top-N approach did not perform as well as the initial result of the edgel method, and performed worse when N increased. This is because, even though the edgel method performed very well, when N was large there were more irrelevant images. If we do not constrain the relevant images calculated using the Top-N approach, many irrelevant images will be falsely selected as relevant and used in the feedback. This will have a negative effect on the final results.

However, the initial results of the ARP method were comparatively low, and many relevant images had low rankings. Therefore, the relevant image grouping from Top-N improved the final performance. Note that in Fig. 12, the performances of Top-50 and Top-100 are identical, so the performance of Top-N is better than the initial result and Top-1+Top-N. But it is lower than the proposed method (Top-3+Top-N). This is because there were more relevant images in the top-R (R>1).

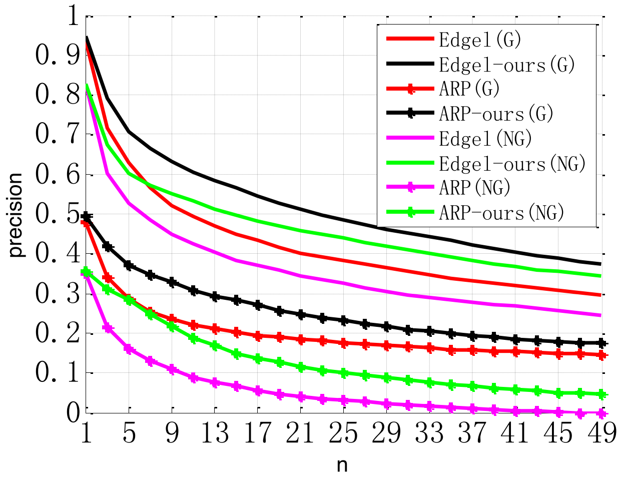


Fig. 14. Comparisons using good (G) and not good (NG) sketches.

The experimental results shown in Fig. 4, Fig. 5, Fig. 11, and Fig. 12 show the effectiveness of our relevant image grouping approach (Top-R+Top-N). Figures 11 and 12 show that for Top-3+Top-N, ours approach performs 7% better than the edgel method, and 13% better than ARP. The performance was approximately 10% better than the other approaches, even when $n = 1$.

We then analyzed the effect of R on the final performance by letting it vary in the range [1,10], and compared our method to ARP. The impact of R on the retrieval performance is shown in Fig. 13. The performance increased when R increased, but the performance decreased when $R > 5$. This is because a larger R introduces more noisy (irrelevant) images to the grouping procedure. However, the result was still better than the initial ARP result.

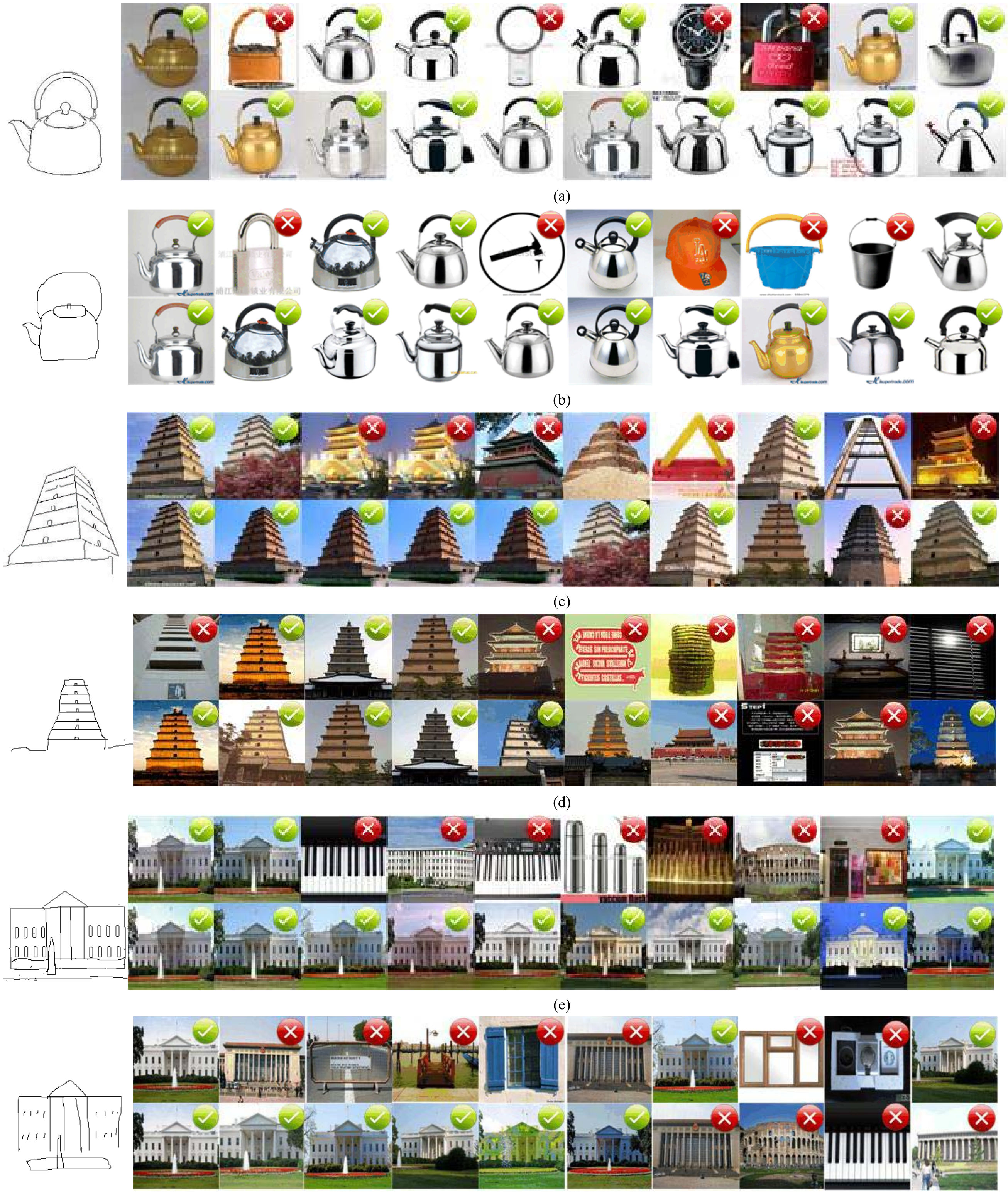
G. Impact of the Sketch Quality

Different users have different drawing skills, so the sketch quality is variable. We systematically evaluated the SBIR performance for different quality sketches using the ARP and edgel based approaches. The results are shown in Fig. 14.

We calculated the precision@n curves for good and bad sketches using our method (Top-1+top-100). In Fig. 14, Edgel(G) and ARP(G) denote the initial approach using good sketches. Edgel(NG) and ARP(NG) denote the initial approach using not good sketches. Edgel-ours and ARP-ours denote the proposed method. Figure 14 shows that bad sketches reduce the performance by approximately 10%, for the edgel and ARP methods. However, regardless of the input, the proposed method performs much better than the other techniques. For the initial edgel system, our method increased the precision by 10%, and for the ARP method, the improvement was approximately 5%.

H. Subjective Comparisons

Figure 15 compares the edgel and proposed methods using good and bad sketches. Figure 15 (a,c,e) shows the retrieval results of the edgel method (the first row) and our method



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 15. Sketch retrieval results using the edgel and proposed methods with three good and three bad sketches. The top rows of each subfigure contain the edgel results, and the bottom rows contain the results of the proposed methods. (a,c,e) are the results for the good sketches, and (b,d,f) are for the bad sketches. (a) Retrieval results of the edgel and proposed methods using a good sketch. (b) Retrieval results of the edgel and proposed methods using a bad sketch. (c) Retrieval results of the edgel and proposed methods using a good sketch. (d) Retrieval results of edgel and proposed methods using a bad sketch. (e) Retrieval results of the edgel and proposed methods using a good sketch. (f) Retrieval results of the edgel and proposed methods using a bad sketch.

(the second row) using good sketches, and (b,d,f) are the corresponding results using bad sketches. When using a good sketch, our top 10 results were all correct and the edgel method returned several irrelevant images. When using a bad sketch, the edgel method returned more irrelevant images, but our top five results were all correct (even though the top edgel result

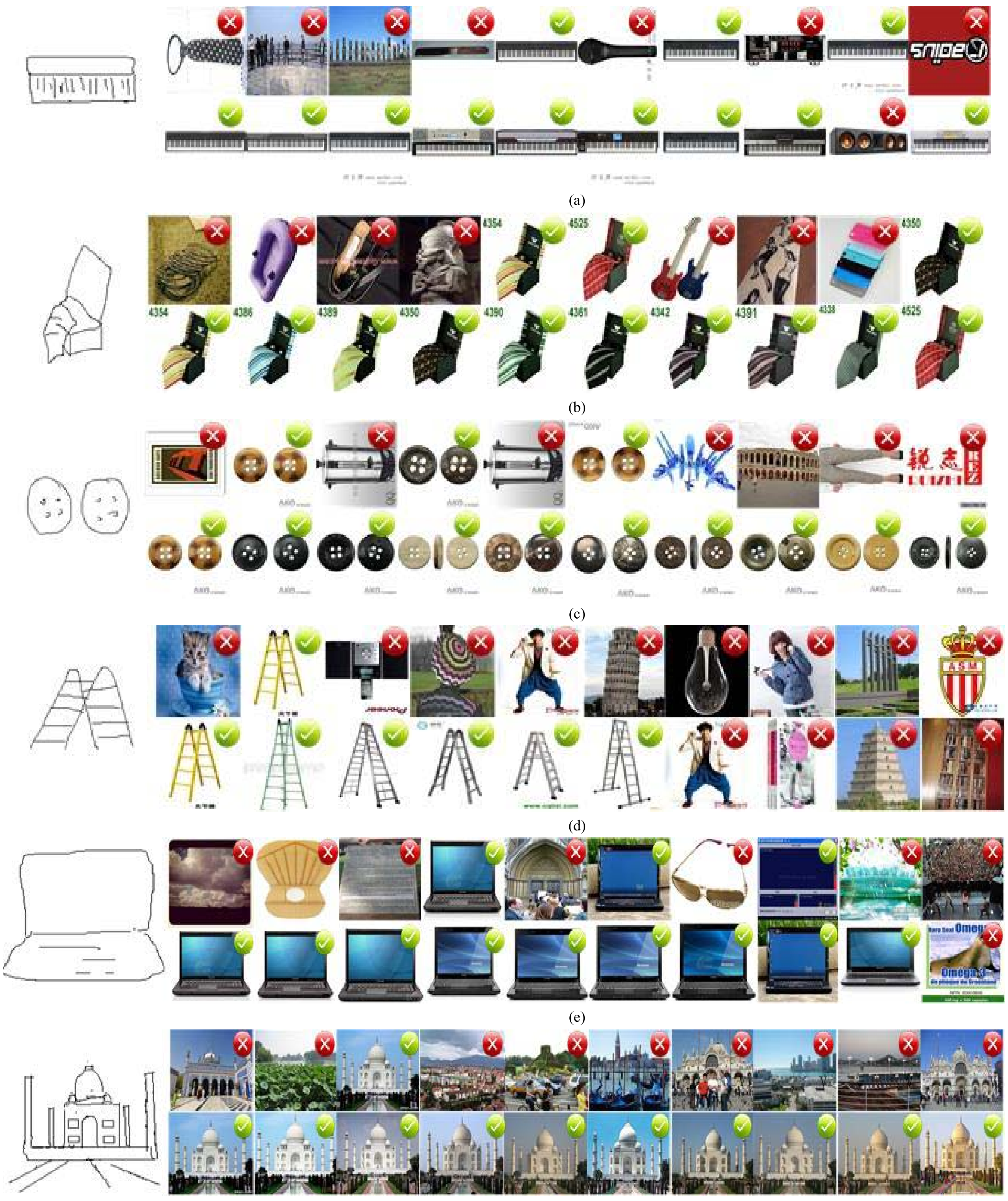


Fig. 16. Sketch retrieval performances, the first row contains the top-ranked results using the ARP method, and the second row contains the results using the proposed method. (a)-(f) are the input query sketches and their corresponding retrieval results of the ARP based approach and our proposed approach.

was an irrelevant image, as shown in Fig. 15 (d)). Figure 15 (e) shows that our results also contained some incorrect images, but they are all similar in shape to the queries, and the results were better than those of the edgel method. There were many irrelevant images that had contours with similar parts to the edgel query sketch. After the RVFV and CBRF stages, these irrelevant images were mostly eliminated.

Figure 16 contains the retrieval results of the ARP and proposed methods for six sketches, where the first SBIR results were not correct. We used the top five results for the relevant image grouping procedure. After the relevant image grouping and content-based relevant feedback stages, more relevant images were found.

VI. Conclusion

We proposed a SBIR method that uses initial result grouping, re-ranking via visual verification, and a relevance feedback system to search for more similar images. The initial result grouping helps our system find more relevant images for the relevance feedback. Our RVFV approach filters out irrelevant images to improve the relevance feedback, and to find more relevant images for the top-ranked images. The proposed CBRF more deeply explores relevant images, to find those that were not found in the original SBIR. These systems work well when compared with other methods, and can find many relevant images when the initial results are sufficient. Note that our approach does not destroy the original index structure, and does not significantly increase time or storage costs. But the proposed method can't find the images with differently size and rotation. In the future work, we will work hard to solve this problem. Theoretically, this method can be combined with a wide range of existing SBIR methods to improve the final retrieval results.

References

[1] A. Chalechale, G. Naghdy, and A. Mertins, "Edge image description using angular radial partitioning," *IEEE Proc.-Vis., Image Signal Process.*, vol. 151, no. 2, pp. 93–101, Apr. 2004.

[2] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 761–768.

[3] E. Di Sciascio, G. Mingolla, and M. Mongiello, "Content-based image retrieval over the Web using query by sketch and relevance feedback," in *Proc. 3rd Int. Conf. VISUAL*, London, U.K., 1999, pp. 123–130.

[4] C. Liu, D. Wang, X. Liu, C. Wang, L. Zhang, and B. Zhang, "Robust semantic sketch based specific image retrieval," in *Proc. IEEE ICME*, Jul. 2010, pp. 30–35.

[5] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, Apr. 2008, Art. ID 5.

[6] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 4, pp. 288–297, 1999.

[7] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, Jan. 2000.

[8] E. Cheng, F. Jing, and L. Zhang, "A unified relevance feedback framework for Web image retrieval," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1350–1357, Jun. 2009.

[9] P. Salembier and F. Marqués, "Region-based representations of image and video: Segmentation tools for multimedia services," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 8, pp. 1147–1169, Dec. 1999.

[10] K. Hirata and T. Kato, "Query by visual example—Content based image retrieval," in *Proc. 3rd Int. Conf. Extending Database Technol., Adv. Database Technol.*, 1992, pp. 56–71.

[11] S. Liang and Z. Sun, "Sketch retrieval and relevance feedback with biased SVM classification," *Pattern Recognit. Lett.*, vol. 29, no. 12, pp. 1733–1741, 2008.

[12] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 5, pp. 489–497, May 1990.

[13] S. Raimondo, G. Ciocca, and I. Gagliardi, "Feature extraction for content-based image retrieval," in *Encyclopedia of Database Systems*. New York, NY, USA: Springer-Verlag, 2009, pp. 1115–1119.

[14] B. Szántó, P. Pozsegovics, Z. Vámossy, and S. Sergyán, "Sketch4match—Content-based image retrieval system using sketches," in *Proc. IEEE 9th SAMI*, Jan. 2011, pp. 183–188.

[15] Y. Wang, M. Yu, Q. Jia, and H. Guo, "Query by sketch: An asymmetric sketch-vs-image retrieval system," in *Proc. 4th Int. Congr. Image Signal Process.*, 2011, pp. 1368–1372.

[16] T. Chen, M. Cheng, P. Tan, A. Shamir, and S. Hu, "Sketch2Photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. ID 124.

[17] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical Bayesian filter," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1372–1384, Sep. 2006.

[18] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 644–655, Sep. 1998.

[19] T.-S. Chua, W.-C. Low, and C.-X. Chu, "Relevance feedback techniques for color-based image retrieval," in *Proc. MMM*, Oct. 1998, pp. 24–31.

[20] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-based image retrieval with relevance feedback in Mars," in *Proc. ICIP*, 1997, pp. 815–818.

[21] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.

[22] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[23] J. Ming, R. Machiraju, and D. Thompson, "Geometric verification of swirling features in flow fields," in *Proc. IEEE Vis.*, Nov. 2002, pp. 307–314.

[24] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 889–896.

[25] M. Okabe and S. Yamada, "Semisupervised query expansion with minimal feedback," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 11, pp. 1585–1589, Nov. 2007.

[26] R. Ullah and J. Jaafar, "Exploiting short query expansion for images retrieval," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, vol. 1. Dec. 2012, pp. 352–356.

[27] G. Akrivas, M. Wallace, G. Andreou, G. Stamou, and S. Kollias, "Context—Sensitive semantic query expansion," in *Proc. ICAIS*, 2002, pp. 109–114.

[28] M. M. Rahman, B. C. Desai, and P. Bhattacharya, "Visual keyword-based image retrieval using latent semantic indexing, correlation-enhanced similarity matching and query expansion in inverted index," in *Proc. 10th IDEAS*, Dec. 2006, pp. 201–208.

[29] J. Philbin, O. Chum, M. Isard, L. Sivic, and A. Zissennan, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. CVPR*, Jun. 2008, pp. 1–8.

[30] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1150–1157.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[32] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. CVPR*, vol. 2. Jun. 2006, pp. 2161–2168.

[33] J. Li, X. Qian, Y. Tang, L. Yang, and T. Mei, "GPS estimation for places of interest from social users' uploaded photos," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2058–2071, Dec. 2013.

[34] J. Li, X. Qian, Y. Y. Tang, L. Yang, and C. Liu, "GPS estimation from users' photos," in *Proc. 19th Int. Conf. MMM*, 2013, pp. 118–129.

[35] N. Y. Khan, B. McCane, and G. Wyvill, "SIFT and SURF performance evaluation against various image deformations on benchmark dataset," in *Proc. Int. Conf. DICTA*, 2011, pp. 501–506.

[36] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han, "Contextual weighting for vocabulary tree based image retrieval," in *Proc. IEEE ICCV*, Nov. 2011, pp. 209–216.

[37] X. Yang, X. Qian, and Y. Xue, "Scalable mobile image retrieval by exploring contextual saliency," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1709–1721, Jun. 2015.

[38] T. Mei, B. Yang, X.-S. Hua, and S. Li, "Contextual video recommendation by multimodal relevance and user feedback," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, 2011, Art. ID 10.

[39] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "multimodal graph-based reranking for Web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2013.

[40] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[41] A. Chalechale, G. Naghdy, and A. Mertins, “Sketch-based image matching using angular partitioning,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 35, no. 1, pp. 28–41, Jan. 2004.

[42] R. H. van Leuken, L. Garcia, X. Olivares, and R. Zwol, “Visual diversification of image search results,” in *Proc. 18th Int. Conf. WWW*, 2009, pp. 341–350.

[43] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen, “Hierarchical clustering of WWW image search results using visual, textual and link information,” in *Proc. 12th ACM MM*, 2004, pp. 952–959.

[44] S. H. Srinivasan and N. Sawant, “Finding near-duplicate images on the Web using fingerprints,” in *Proc. 16th ACM MM*, 2008, pp. 881–884.

[45] W.-L. Zhao and C.-W. Ngo, “Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection,” *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 412–423, Feb. 2009.

[46] S. Zhang, Q. Tian, K. Lu, Q. Huang, and W. Gao, “Edge-SIFT: Discriminative binary descriptor for scalable partial-duplicate mobile search,” *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2889–2902, Jul. 2013.

[47] X. Qian, Y. Zhao, and J. Han, “Image location estimation by salient region matching,” *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 4348–4358, Nov. 2015.

[48] X. Qian, Y. Zhang, and X. Tan, “Sketch-based image retrieval using contour segments,” in *Proc. IEEE MMSP*, pp. 1–6, Oct. 2015.

Xueming Qian (M’10) received the B.S. and M.S. degrees from the Xi’an University of Technology, Xi’an, China, in 1999 and 2004, respectively, and the Ph.D. degree from the School of Electronics and Information Engineering, Xi’an Jiaotong University, in 2008. He was a Visiting Scholar with Microsoft Research Asia from 2010 to 2011. He was an Assistant Professor with Xi’an Jiaotong University, where he was an Associate Professor from 2011 to 2014, and is currently a Full Professor. He is also the Director of the Smiles Laboratory at Xi’an Jiaotong University. He received the Microsoft Fellowship in 2006. He received outstanding doctoral dissertations of Xi’an Jiaotong University and Shaanxi Province, in 2010 and 2011, respectively. His research interests include social media big data mining and search. His research is supported by the National Natural Science Foundation of China, Microsoft Research, and Ministry of Science and Technology.

Xianglong Tan is currently pursuing the M.S. degree with the Smiles Laboratory, Xi’an Jiaotong University, Xi’an, China. His research interests include large scale sketch-based image retrieval and image content understanding.

Yuting Zhang is currently pursuing the M.S. degree with the Smiles Laboratory, Xi’an Jiaotong University, Xi’an, China. Her research interests include large scale sketch-based image retrieval and image content understanding.

Richang Hong (M’12) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Fellow with the School of Computing, National University of Singapore, from 2008 to 2010. He is currently a Professor with the Hefei University of Technology, Hefei. He has co-authored over 60 publications in the areas of his research interests, which include multimedia question answering, video content analysis, and pattern recognition. He is a member of the Association for Computing Machinery. He was a recipient of the best paper award in the ACM Multimedia 2010.

Meng Wang (M’09) received the B.E. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China. He is currently a Professor with the Hefei University of Technology, China. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing. He received the special class for the Gifted Young from USTC. He received the best paper awards successively from the 17th and 18th ACM International Conference on Multimedia, the best paper award from the 16th International Multimedia Modeling Conference, the best paper award from the Fourth International Conference on Internet Multimedia Computing and Service, and the best demo award from the 20th ACM International Conference on Multimedia.